Studies in Data Annotation Effectiveness

Michael R. Crystal, Francis Kubala, Robert MacIntyre

BBN Technologies

Cambridge, MA 02138-1119

{mcrystal,fkubala,rmacinty}@bbn.com

Introduction

BBN has successfully applied the train-by-example paradigm to both speech and natural language understanding problems. Per this paradigm, systems are created not by experts encoding their knowledge in a form such as rules but, rather by being given examples of correct behavior, creating a model from those examples, and then applying the model to novel situations.

Problems to which we have successfully applied this paradigm include speech recognition, speaker identification, topic identification, named entity extraction, and relationship extraction. As a result of this success, we are expanding our research to more diverse and difficult problems such as scenario extraction from text and multilingual information extraction.

As we develop more train-by-example systems, we have an increasing need for highly effective, low-cost training data - the examples. In most cases, preparing training data is a laborious task conducted by humans. The study related herein quantifies the effectiveness and cost of different approaches to annotating data for our named-entity extraction from speech system.

Experimental Design

For this study, we considered the effects of two annotation parameters on extraction F-scores: 1) Annotator Experience and 2) Annotation Quality. During the study, we recorded the amount of time annotators spent achieving different levels of annotation quality. Based on these times, we evaluated the cost of achieving extraction quality.

We began with a baseline training set of documents that included roughly 25K Named-Entities (NEs). We then augmented the baseline with a study set. The study set, comprised of roughly 36K NEs, was annotated with six different approaches. The different approaches varied annotator experience and annotation quality. Finally, we measured the effectiveness (F-score) of the models based on the augmented training sets.

The study involved six annotators. We refer to them as R, A, G, E, C, and H. Table 1 lists the experience and educational background of each annotator. The three seasoned annotators, all college graduates, had been performing named entity annotation for at least ten weeks each. Our novice annotators had one or two weeks of named entity annotation training prior to beginning work on the study; otherwise, they had never performed any annotation task.

We characterize annotation quality (and effort) by the combination of atomic annotation steps applied to a document set. The four atomic processing steps are listed in Table 2.

Code	Experience	Education
R	>10 weeks	College graduate, non-technical field
A	>10 weeks	College graduate, non-technical field
G	>10 weeks	College graduate, non-technical field
E	6 weeks	College graduate, technical field
C	2 weeks	In Ph.D. program for linguistics
H	1 week	College sophomore

Table 1 Annotator Experience

Human Annotation:	A person annotates a document.
Machine Annotation:	A machine annotates a document based on an existing model.
Human Adjudication:	A person adjudicates discrepancies between any two annotated versions of a document.
Test-on-Train:	A three-step process: 1. A named entity model is trained with an annotated document set. 2. Based on the model, machine annotation is performed on the training set. 3. A human adjudicates between the training set and the machine annotation results.

Table 2 Atomic Annotation Processing Steps

Based on the four atomic processing steps, we produced four composite quality levels. We refer to the first as *single-annotation*. Human annotation was performed once on a document set and the document set was used to create an extraction model.

The second quality level is *single, test-on-train*. In this case, the result of a single annotator's work is run through a test-on-train cycle. In general, the test-on-train annotation agrees with the annotator in 99% of the cases. The remaining 1% draw attention to inconsistent annotation by the annotator. Roughly one-quarter of the disagreements are due to human annotator error. The test-on-train cycle also identifies low quality annotation. If test-on-train annotation diverges from the original annotation in more than 2% of the cases (i.e. an F-score less than 98), it is worth reviewing the annotator's work.

The third quality level is *double-annotation*; two human annotators annotate each document and a third annotator adjudicates between the inter-annotator discrepancies. It is important to note that the adjudicator is limited to choosing between one of the two annotators' mark-ups. The adjudicator does not review sections of the documents for which the annotators agreed.

The highest quality level we considered for this study was to perform a *test-on-train cycle based on double-annotated* materials. If we accept the plausible claim that an adjudicator does not increase error rate, then we are guaranteed that adjudication and test-on-train monotonically increase the quality of annotation.

We began with roughly 1.1 million words of Hub-4 transcribed broadcast news. We divided our document set into three partitions. The approximate word counts and named entity (NE) counts for each partition are listed in Table 3.

Annotation requirements for training models are not uniform. Initially bootstrapping a system, moving performance from inadequate to sub-optimal, and optimizing performance require different quality annotation. This study is concerned with annotation that increases performance but does not necessarily optimize it. So that we could focus on annotating to improve performance, we began the study by having our most experienced annotators and adjudicators perform the highest quality annotation on partitions 1 and 2. Specifically, A and G double-annotated partition-1 and partition-2. R and B adjudicated and performed a test-on-train cycle on the two partitions. These two partitions defined our baseline.

Training on partition-1 yielded a model with an F-score of 81.3 with respect to a blind test set that is used for all testing. Training on partition-2 yielded a model with an F-score of 81.5. A model trained on both partitions yielded a score of 83.0.

Partition	Document Count	Word Count	NE Count
1	44	246,581	13,021
2	44	218,489	11,844
3	98	621,795	35,734
Total	**186**	**1,102,196**	**60,599**

Table 3 Annotation Quantity

We performed the core of the study by creating six, differently annotated versions of partition-3. We compared the baseline performance to the performance of models created by augmenting the baseline data with each partition-3 version - set-1 through set-6. Figure displays the process by which the six annotated were created.

During this portion of the experiment, we ran twenty-one experiments. We began with partition-1, created a model with it, and evaluated the model. We then added, in turn, each of the six annotation sets to partition-1, created a model, and evaluated the model. Next we repeated the seven experiments with partition-2 and, then, we repeated the seven experiments starting with a combination of both partitions 1 and 2. The F-scores that resulted from evaluating the twenty-one models are shown in Table 4 and its associated figure.

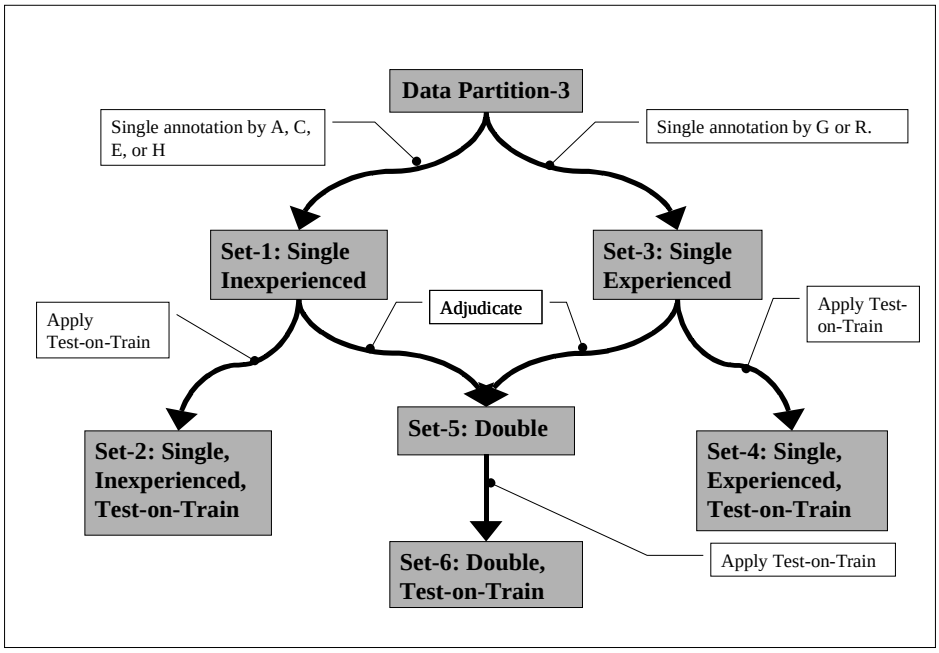
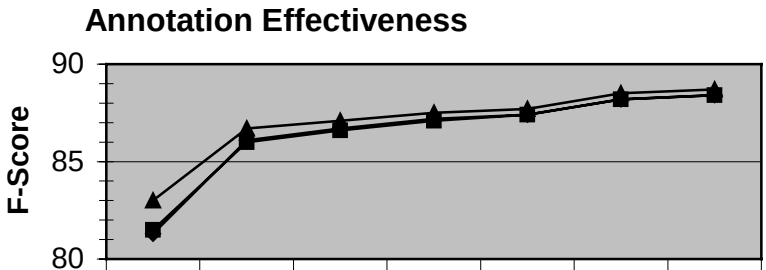


Figure 1 Anotation Process



	Base	Set-1	Set-2	Set-3	Set-4	Set-5	Set-6
Partition-1	81.3	86.1	86.7	87.2	87.4	88.2	88.4
Partition-2	81.5	86.0	86.6	87.1	87.4	88.2	88.4
Partitions 1 & 2	83.0	86.7	87.1	87.5	87.7	88.5	88.7

Annotation Approach

Table 4 Annotation Effectiveness

ANNOTATOR	# Documents Annotated	SCORE	Test-on-Train SCORE
R	14	98.52	98.90
G	100	96.72	97.28
A	49	95.94	96.56
C	22	95.44	96.65
E	22	93.78	96.16
H	21	91.97	94.35
AVERAGE	38	95.40	96.65
Adjudicated		99.47	100 (by definition)

Table 5 Annotator Quality

To assess the effects of annotator experience on model quality, we teased apart sets 1, 2, 3, and 4 and re-batched them by annotator. This resulted in six batches of single annotated materials from sets 1 and 3 - one each for R, G, A, C, E, and H. and another six test-on-train batches from sets 2 and 4. We then compared the batches to set-6, the highest quality annotation we had for any particular document; we did not create extraction models for this experiment. The F-scores that resulted from this comparison are shown in Table 5.

To assess the cost of each of the annotation processing steps and cost incurred by using annotators with different experience levels, we recording the time required by each annotator to create sets 1 through 6. Although annotators do not have uniform costs associated with them, we assume that they do for purposes of this study. **Error! Reference source not found.** lists processing time by annotator and task.

Analysis

For this study we considered three types of data refinement and their interactions:

1. The number of times the data was annotated - single or double annotation
2. The experience of the annotators
3. Whether the data was put through a quality assurance (test-on-train) cycle.

Data sets 1 and 3 in Figure 1 are single annotated training sets, set 1 by inexperienced staff and set 3 by experienced staff. Adding set 1 to our base training generated a roughly 4.3 point increase in F-score. Adding set 3, our F-score increased by 5.3 points. When we add set-5, the results of adjudicating sets 1 and 2, we gain 6.4 points over the base. This performance increase is 50% greater than that of inexperienced annotators and 21% greater than that of experienced annotators.

Data sets 1 and 2 were created by inexperienced annotators. Data sets 3 and 4 are the same documents annotated by experienced annotators. Models resulting from single annotation by experienced annotators increased F-scores by 23% (+5.3 vs. +4.3) more than by their inexperienced counterparts. If the models are based on annotation that has been passed through a test-on-train QA phase, then the F-score increase is reduced to 17% (+5.6 vs. +4.9).

Annotator hours	***P1&P2***	*Set-1*	*Set-2*	*Set-3*	*Set-4*	*Set-5*	*Set-6*	***TOTAL***
R				10				10
A	48	5						53
G	79			10				89
E		12						12
C		31	33		33	18	32	147
H		38						38
B	24					11		35
TOTAL	151	86	33	20	33	29	32	384

Table 6 Annotation Cost

Comparing data sets 1 to 2, 3 to 4, and 5 to 6 illustrates the effects of applying a test-on-train QA cycle. Models based on inexperienced single annotation are 12%(+4.9 vs. +4.3) better when passed through a test-on-train cycle. When test-on-train is applied to experienced annotators' work, the net effect is 4% increase greater increase in F-score (+5.6 vs. +5.3). When applied to doubly annotated training data, the net effect is only a 3% greater F-score increase (+6.6 vs. +6.4).

One can conclude from Table 4 and these analyses that the most important factor for increasing annotation effectiveness, that is the F-score of the resulting model when applied to a test set, is simply creating annotated data. Further gains can be made by first, using experienced annotators. Experience here refers to annotation experience not computer science or linguistics background. Second, gains are achieved by double-annotating data; and, third, one can perform test-on-train quality assurance.

Furthermore, one can conclude that these approaches, experience, number of annotation passes, and QA passes are not linearly independent. Experienced annotators have less to gain from double annotation and QA. QA is more useful when applied to singly annotated materials or those produced by inexperienced annotators.

Error! Reference source not found., gives an approximate cost of each of the three aforementioned processing steps. For simplicity sake, we consider the cost of all annotators to be equal. Experienced annotators worked at roughly 4x the speed of the inexperienced annotators. This means that in addition to increasing performance by between 17% and 23%, experience decreases cost markedly.

Double-annotation costs are dominated by the cost of the slower annotators. Hence, moving from experienced annotation to double-annotation raises the cost by a factor of 7.4, while it increases the cost of inexperienced annotation by a factor of only 1.7. For this study, we did not consider the cost of double-annotation by like-experienced annotators - we assumed the factor would be between 2 and 3 times.

Although we concluded that annotation quantity is the most important factor for increasing model performance, we believe that several implicit aspects of our study contribute to this conclusion. First, we began with sub-optimal, but nonetheless trained models based on partitions 1 and/or 2. We believe that, initial model training requires much higher quality data than training used to incrementally increase performance. In the same vein, we were trying to achieve increases starting with F-scores in the low 80's. If we were to try to increase performance from the mid 90's, we believe that we would also require very high quality data that could only be generated by passing it through all refinement steps.

We also believe that annotation quality is less of a factor when the annotation process includes well-established protocols. Each of our annotators was given detailed, written guidelines at the project outset. These guidelines included examples of how to handle most ambiguities. An annotation effort that does not include these protocols will need to rely more heavily on adjudication and test-on-training cycles to ferret out training shortcomings, guideline ambiguities, or other problems that can lead to lower quality annotation.

It is also worth pointing out that all of our annotators are native English speakers with college degrees. Our study does not establish a lower bound on annotator experience or annotation quality requirements below which performance will be adversely effected; however, we believe that non-native speakers of a language would cross this threshold.

Based on these results, we argue that when annotation resources are scarce and F-Scores are below 90 (which applies to the text-from-speech domain), model training should employ experienced annotators to focus on creating high quantity training examples, foregoing quality assurance at the expense of decreased training quality.

Further Directions

Given our conclusions about the importance of annotator experience, further research should include methods for selecting and training annotation staff. As we continue to expand our annotation efforts, we have begun this work. Earlier this year we published a protocol for establishing annotation teams.

This study focused on a subset of annotation techniques applied to one phase of the model-training problem. Further research is necessary if we are to draw any conclusions about annotation for initial model creation or optimization for high performance applications.

We also need to further consider the tradeoffs between annotation quality and quantity. Would increasing the error rate by 2 and annotation rate by 4 still result in increased F-Scores for a fixed amount of annotation dollars?

Other annotation techniques that we have applied at BBN, but did not consider in study include:

- Double annotating, but having the machine be one of the annotators.

- Having the annotator correct machine annotations instead of starting from scratch.
- Adding an *unsure* tag so that annotators can quickly skip over, and come back to time-consuming ambiguities.
- Performing keyword-in-context analysis on words that are often annotated incorrectly.
- Dumping all annotated entities to a file for a *sanity check*.
- Preprocessing training materials so that we are annotating the materials that will do the most to extend model performance.

Acknowledgements

This work was supported by the Air Force Research Laboratory under contract number F30602-97-C-0253. The views and findings contained in this material are those of the authors and do not necessarily reflect the position or policy of the U.S. Government and no official endorsement should be inferred.